Mining Clickthrough Data to Improve Search Engine Results

Ashok Veilumuthu (Sr.No: 4-04-05-1-04242-0)
PhD Student
Department of Management Studies

Abstract

In this thesis, we aim at improving the search result quality by utilizing the search intelligence (history of searches) available in the form of clickthrough data. We address two key issues, namely 1) relevance feedback extraction and fusion, and 2) deciphering search query intentions.

Relevance Feedback Extraction and Fusion: The existing search engines depend heavily on the web linkage structure in the form of hyperlinks to determine the relevance and importance of the documents. But these are collective judgments given by the page authors and hence, prone to collaborated spamming. To overcome the spamming attempts and language semantic issues, it is also important to incorporate the user feedback on the documents' relevance. Since users can be hardly motivated to give explicit/direct feedback on search quality, it becomes necessary to consider implicit feedback that can be collected from search engine logs. Though a number of implicit feedback measures have been proposed in the literature, we have not been able to identify studies that aggregate those feedbacks in a meaningful way to get a final ranking of documents.

In this thesis, we first evaluate two implicit feedback measures namely 1) click sequence and 2) time spent on the document for their content uniqueness. We develop a mathematical programming model to collate the feedbacks collected from different sessions into a single ranking of documents. We use Kendall's τ rank correlation to determine the uniqueness of the information content present in the individual feedbacks. The experimental evaluation on top 30 select queries from an actual search log data confirms that these two measures are not in perfect agreement and hence, incremental information can potentially be derived from them. Next, we study the feedback fusion problem in which the user feedbacks from various sessions need to be combined meaningfully.

Preference aggregation is a classical problem in economics and we study a variation of it where the rankers, i.e., the feedbacks, possess different expertise. We extend the generalized Mallows' model to model the feedback rankings given in user sessions. We propose a single stage and two stage aggregation framework to combine different feedbacks into one final ranking by taking their respective expertise into consideration. We show that the complexity of the parameter estimation problem is exponential in number of documents and queries. We develop two scalable heuristics namely, 1) a greedy algorithm, and 2) a weight based heuristic, that can closely approximate the solution. We also establish the goodness of fit of the model by testing it on actual log data through log-likelihood ratio test. As the independent evaluation of documents is not available, we conduct experiments on synthetic datasets devised appropriately to examine the various merits of the heuristics. The experimental results confirm the possibility of expertise oriented aggregation of feedbacks by producing orderings better than both the best ranker as well as equi-weight aggregator. Motivated with this result, we extend the aggregation framework to hold infinite rankings for the meta-search applications. The aggregation results on synthetic datasets are found to be ensuring the extension fruitful and scalable.

Deciphering Search Query Intentions: The search engine often retrieves a huge list of documents based on their relevance scores for a given query. Such a presentation strategy may work if the submitted query is very specific, homogeneous and unambiguous. But many a times it so happen that the queries posed to the search engine are too short to be specific and hence ambiguous to identify clearly the exact information need, (eg. "jaguar"). These ambiguous and heterogeneous queries invite results from diverse topics. In such cases, the users may have to sift through the entire list to find their needed information and that could be a difficult task. Such a task can be simplified by organizing the search results under meaningful subtopics, which would help the users to directly move on to their topic of interest and ignore the rest.

We develop a method to determine the various possible intentions of a given short generic and ambiguous query using information from the click-through data. We propose a two stage clustering framework to co-cluster the queries and documents into intentions that can readily be presented whenever it is demanded. For this problem, we adapt the spectral bipartite partitioning by extending it to automatically determine the number of clusters hidden in the log data. The algorithm has been tested on selected ambiguous queries and the results demonstrate the ability of the algorithm in distinguishing among the user intentions.